Streamed Databases

The definition of a streaming database is a data store that is designed to process, collect and enrich a series of incoming data points in real-time (Hazelcast, 2021). A streaming database is different from a traditional RDBMS in terms of how the database administrator uploads data, it will typically be uploaded via an ETL process or tool at even intervals.

A streamed database can be used alongside a RDBMS, to use it for modern cases or in larger enterprises (Hazelcast, 2021). Data needs to be processed incrementally and sequentially and can be used for analytics such as filtering, correlations, sampling, or aggregations (AWS, 2021). The use of streamed databases has increased because of how continuously data accelerates. Technologies that use to rely primarily on batch-orientated, switched to relying on more heavy streaming database technologies (Hazelcast, 2021).

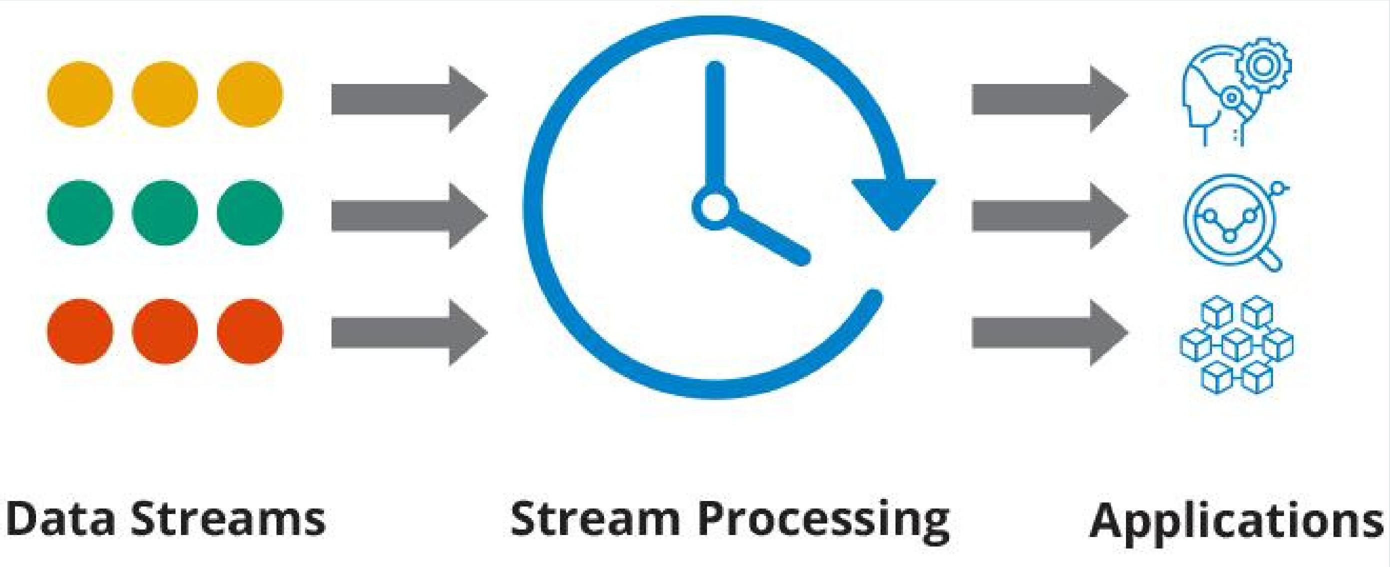
Data Architecture

Streaming data architecture is a framework of software components, used for processing large amounts of streaming data that comes from different sources (Levy, 2021). A streaming data architecture takes in data immediately as it is generated, whereas traditional data solutions read and write data in batches (Levy, 2021).

Benefits of Streaming Data Architecture

- Streaming Databases are able to deal with never-ending streams of events and give insight into large amounts of data.
- Real-time or near-real-time processing, this is useful for giving feedback on how situations are at that specific moment.
- Detecting patterns in time-series data, for example looking for trends in website data.
- Easy data scalability, growing data can break traditional systems, but streaming databases are hyperscalable

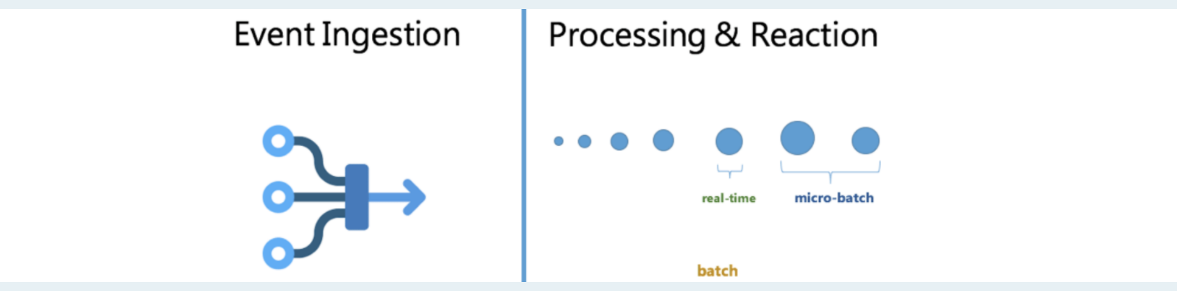
Collecting data in a streaming database



Kafka

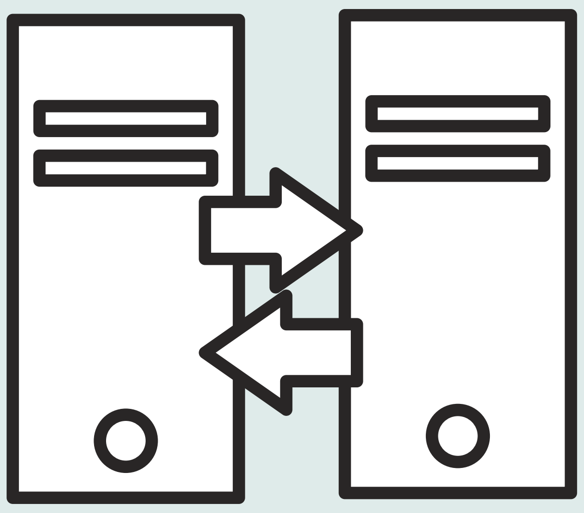
Kafka is used to streaming data in real-time from heterogeneous sources such as SQLServer or MySQL (Tortella, 2020). Kafka has the ability to create topics based on the objects from these sources in order to stream data in real-time.

This data can be used to visualize real-time data with the use of visualization tools or to populate any destination system (Tortella, 2020).



Types of Data:

- Log Files
- Information from social networks
- eCommerce purchases
- Financial trading floors
- In-game player activity
- Telemetry



Streaming Dataflow

In traditional batch solutions, data has to be ingested, processed, and then structured before it can be used, streaming data has the ability to consume, persist and analyze data while it is in motion (Confluent, 2020).

Applications working with streamed data will always require two functions: storage and processing (Confluent, 2020). Storage is required to store large amounts of data in a way that is both consistent and sequential. The processing has to be able to interact with the storage, analyze, and run services on the consumed data (Confluent, 2020).

Useful applications

Examples of streaming data Analytics Tools:

- Amazon Athena
- Amazon Redshift
- Elasticsearch
- Cassandra

Examples of modern streaming architectures:

- Real-time Machine Learning at Bigabid
- Multi-purpose data lake at ironSource
- Transitioning from the data warehouse to data lake at Meta Networks
- Automation of data plumbing

Made By: Enrico Dreyer (31210783)

ITRI 623 Assignment 2